Effects of Glasgow Outcome Scale misclassification on traumatic brain injury clinical trials

JOURNAL OF NEUROTRAUMA 25:641–651 (June 2008)

DOI: 10.1089/neu.2007.0510

Effects of Glasgow Outcome Scale Misclassification on Traumatic Brain Injury Clinical Trials

JUAN LU,[1] GORDON D. MURRAY,[4] EWOUT W. STEYERBERG,[3] ISABELLA BUTCHER,[4] GILLIAN S. MCHUGH,[4] HESTER LINGSMA,[3] NINO MUSHKUDIANI,[3] SUNG CHOI,[2] ANDREW I.R. MAAS,[5] and ANTHONY MARMAROU[1]

ABSTRACT

The Glasgow Outcome Scale (GOS) is the primary endpoint for efficacy analysis of clinical trials in traumatic brain injury (TBI). Accurate and consistent assessment of outcome after TBI is essential to the evaluation of treatment results, particularly in the context of multicenter studies and trials. The inconsistent measurement or interobserver variation on GOS outcome, or for that matter, on any outcome scales, may adversely affect the sensitivity to detect treatment effects in clinical trial. The objective of this study is to examine effects of nondifferential misclassification of the widely used five-category GOS outcome scale and in particular to assess the impact of this misclassification on detecting a treatment effect and statistical power. We followed two approaches. First, outcome differences were analyzed before and after correction for misclassification using a dataset of 860 patients with severe brain injury randomly sampled from two TBI trials with known differences in outcome. Second, the effects of misclassification on outcome distribution and statistical power were analyzed in simulation studies on a hypothetical 800-patient dataset. Three potential patterns of nondifferential misclassification (random, upward and downward) on the dichotomous GOS outcome were analyzed, and the power of finding treatments differences was investigated in detail. All three patterns of misclassification reduce the power of detecting the true treatment effect and therefore lead to a reduced estimation of the true efficacy. The magnitude of such influence not only depends on the size of the misclassification, but also on the magnitude of the treatment effect. In conclusion, nondifferential misclassification directly reduces the power of finding the true treatment effect. An awareness of this procedural error and methods to reduce misclassification should be incorporated in TBI clinical trials.

Key words: clinical trial; Glasgow Outcome Scale; misclassifications; observer variation; power; traumatic brain injury

Departments of [1]Neurosurgery and [2]Biostatistics, Virginia Commonwealth University, Richmond, Virginia.
[3]Center for Medical Decision Sciences, Department of Public Health, Erasmus Medical Center, Rotterdam, The Netherlands.
[4]Department of Community Health Sciences, University of Edinburgh, Scotland.
[5]Department of Neurosurgery, University Hospital, Antwerp, Belgium.

INTRODUCTION

Accurate and consistent assessment of outcome after traumatic brain injury (TBI) is essential to the evaluation of treatment results, particularly in the context of multicenter studies and trials. Various studies have investigated inter-observer agreement and misclassification of TBI outcome measures commonly used in TBI studies, and in general found that interobserver variation or misclassification on Glasgow Outcome Scale (GOS) outcome does exist (Anderson et al., 1993; Brooks et al., 1986; Choi et al., 2002; Maas et al., 1983; Marmarou, 2001; Pettigrew et al., 2003; Scheibel et al., 1998; Teasdale et al., 1998; Wilson et al., 1998, 2002, 2007), ranging from 17% (Marmarou, 2001) to 40% (Wilson et al., 2007) in practices. Previous work has shown that this could attenuate the true treatment effect and reduces the power of detecting the efficacy of treatment (Choi et al., 2002). However, little is known on how different misclassification directions or patterns might affect analysis of treatment effects in double-blinded TBI trials. It would seem reasonable to suspect that misclassification in a clinical trial would possibly effect both the treatment and control groups equally. However, even in this case, there is a profound effect on the analysis of the treatment effects (Choi et al., 2002). In clinical practice, nondifferential misclassification may affect the GOS outcome through three potential patterns. The *random* pattern refers to the misclassification between the adjacent categories that have an equal rate or chance of being classified for both treatment groups. The *upward* pattern means more true outcome categories are classified into better outcome categories for both groups. The *downward* pattern means more true outcome categories are classified into less optimistic outcome categories for both groups. The objective of this study is to investigate whether these three potential patterns of measurement error may have differential effects on the power of finding treatment differences in double blinded TBI trials.

METHODS

Misclassification

Misclassification in this paper is defined as an incorrect classification of the GOS outcome in TBI trials. Furthermore, for the purpose of discussing the outcome analysis of a double-blinded TBI trial, we assume in this study that the rates of misclassification are the same for both treatment and control groups. Thus, the outcome misclassification discussed in this study is nondifferential or random, and as defined above includes three potential patterns: (1) random, (2) upward, and (3) downward for both treated and control groups. Realizing that misclassification may be a combination of upward and downward grading in either the placebo or treatment group, we selected patterns, which combined both directions of misclassifications. More specifically, we defined the "upward" pattern as 20% of patients in both control and treated groups misclassified to a higher outcome category and 10% of patients misclassified in a lower outcome category. The downward pattern was defined as 20% of patients misclassified in a lower outcome category, and 10% of patients misclassified to a higher category. These hypothetical percentages of misclassification are in the range of GOS misclassification found in other studies (Anderson et al., 1993; Maas et al., 1983; Marmarou, 2001; Wilson et al., 1998). Our focus in this report is to study misclassification applied equally to placebo and treated groups. However, an imbalance or non-random misclassification among treated and control groups is not considered in this report. Among all five categories of GOS outcome [Death (D), Vegetative (V), Severe Disability (SD), Moderate Disability (MD), and Good Recovery (GR)], only the category of death can be excluded from misclassification, whereas the other four categories are all subject to misclassification, albeit to a different degree. To study the effect of misclassification, it is assumed that a certain rate of misclassification exists in a patient's outcome in two adjacent categories.

Patient Data

For analysis of the effect of misclassification on the outcome differences, we used a dataset of 860 patients with severe brain injury randomly sampled from two TBI trials with known differences in outcome (Hukkelhoven et al., 2002).

For a more detailed analysis of the effect of misclassification on outcome distribution and statistical power, we used a hypothetical 800-patient dataset (400 patients in each arm). In this dataset, a 55% favorable outcome and a 20% mortality outcome distribution was considered as baseline. For both approaches, the GOS was dichotomized into favorable (GR/MD) versus unfavorable (SD/V/D).

Statistical Analysis

Three patterns of misclassification on dichotomized GOS were studied: (1) random pattern, where 20% GOS outcomes were equally misclassified between favorable and unfavorable outcome categories for both study groups; (2) upward pattern, where 20% unfavorable outcomes were misclassified into favorable, and 10% favorable into unfavorable for both study groups; and (3) downward pattern, where 20% favorable outcomes were

misclassified into unfavorable, and 10% unfavorable into favorable for both study groups. For a dichotomous GOS outcome (GR/MD vs. SD/V/D), the simulated misclassification rates were only applied among survivors (i.e., between GR/MD and SD/V); however, all outcomes, including death, were assessed in the final outcome distribution measurement [i.e., (number of favorable outcomes/treatment total)–(number of favorable outcomes/control total).

Power Calculation

In this study, the power was defined as the probability of finding the difference between the treatment and control groups with a 95% two-sided significance. The calculation was based on a range of hypothetical two-proportion comparisons. No covariates were considered to simplify the problem. The powers, under a hypothetical condition with no misclassification and three simulated cases with misclassification, were compared.

The treatment effect in the hypothetical dataset was created following a conventional method (Bolland et al., 1998). For example, 10% treatment effect on a dichotomous GOS outcome [favorable (GR/MD) vs. unfavorable (SD/V/D)] was defined as an overall 10% outcome shift from the unfavorable to the favorable outcome in the treatment group; i.e., the favorable outcome in the treatment group increased by 10%, and the unfavorable outcome decreased by 10% from the baseline.

Further, recognizing no misclassification on the outcome of death within the unfavorable outcome category, we applied the hypothetical treatment effect into the outcome of death, and the remaining unfavorable outcomes (i.e., a combined SD and V) individually. For the outcome of death, 10% treatment effect was defined as a 10% absolute reduction of the baseline numbers, and it was assumed that 10% of patients' outcomes were improved from death to better outcome categories including V and SD. Finally, the remaining numbers of unfavorable outcomes (SD/V) equaled the total treatment group minus 10% of the increased baseline favorable outcome numbers and minus 10% deducted baseline death numbers. The two-sided Chi-Square test was used for the dichotomous outcome comparisons.

RESULTS

Effect of Misclassification

The effects of misclassification on the dichotomous outcome estimation were demonstrated by an actual phase III TBI trial patient dataset displayed in Table 1. It was assumed that there were certain rates of outcome

Table 1. Effect of Misclassifications on the Observed Dichotomous GOS Outcomes

		GOS after misclassification corrections[b]											
		Observed dichotomous GOS[a]			*Random: 20% up and 20% down*			*Upward: 20% up and 10% down*			*Downward: 10% up and 20% down*		
		Unfav.		*Fav.*	*Unfav.*		*Fav.*	*Unfav.*		*Fav.*	*Unfav.*		*Fav.*
Groups	N	*D*	*V/SD*	*MD/G*	*D*	*V/SD*	*MD/G*	*D*	*V/SD*	*MD/G*	*D*	*V/SD*	*MD/G*
Treatment	430	93	85	252	93	29	308	93	73	264	93	25	312
Control	430	131	81	218	131	35	264	131	73	226	131	30	269
Difference (%)[c]			7.9			10.2			8.8			10.0	
P-value[d]			0.020			0.002			0.009			0.002	

[a]Observed Glasgow Outcome Scale: D, death; V, vegetative; SD, severe disabled; MD, moderate disabled; G, good recovery.

[b]The corrected GOS misclassifications are given by the equation: Fav(Observed) = Fav(True) − Rat1*Fav(True) + Rate2*[N − D − Fav(True)]. Where 1) Fav(Observed) is the count of the observed favorable outcomes, 2) Fav(True) is the count of the corrected favorable outcomes, 3) Rate 1 and Rate 2 are the rates of upward and downward misclassification respectively, and 4) N and D represent the group total and the number of deaths respectively. For example, after correcting for 20% upward and 20% downward misclassification, the equation 252 = X − 0.2*X + 0.2*(430 − 93 − X)_ gives the corrected MD/G = 308, and V/SD = 430 − 93 − 308 = 29 for the treatment group; while equation 218 = X − 0.2*X + 0.2*(430 − 131 − X) gives corrected MD/G = 264 ad V/SD = 430 − 131 − 264 = 35 for the control group.

[c]Difference (%) in the favorable outcomes between the treatment and control groups.

[d]Chi-Square Test (two-sided).

categories being misclassified. Thus, reversing the hypothetical misclassified outcome numbers to the observed outcome data would be helpful in gauging the effect of misclassification on the outcome analysis, and the three possible misclassification models were applied.

Random Pattern

In the random pattern, the adjacent outcome categories have an equal rate of being misclassified for both treatment and control groups. For example, in Table 1, it was assumed that equal rates (20%) of patients had been misclassified as favorable or unfavorable outcome for both groups. If these misclassified outcome numbers were corrected based on our assumptions, the true underlying number of patients with the favorable outcomes would be 308 for the treatment group, 264 for the control, and the percentage difference in favorable outcomes between the two groups would be (308 – 264)/430 or 10.2% (p-value = 0.002). The method for calculation is shown in the Table 1 legend. Before the 20% misclassification correction, the observed difference is 7.9 and p-value is 0.02. Thus, misclassification introduces an error of 2.3% (10.2–7.9).

Upward Pattern

The upward model resulted in an upward trend of misclassification for both treatment and control groups, where the rate of patients being misclassified was higher (20%) from the unfavorable outcomes to the favorable outcomes than the rate exchange from the other direction (10%). If the misclassified outcome numbers were corrected, the number of patients with the favorable outcomes would be 264 for the treatment group, and 226 for the control. The actual percentage difference in favorable outcomes between the two groups would be 8.8 (p-value = 0.009) instead of the observed difference of 7.9 (p-value = 0.02). In this case, misclassification introduces an error of 0.9% (8.8–7.9).

Downward Pattern

In the downward model, the rate of being misclassified was lower (10%) from the unfavorable outcomes to the favorable outcomes than the rate exchange from the other direction (20%) for both treatment and control groups. After the misclassified outcome numbers were corrected, the number of patients with the favorable outcome would be 312 for the treatment group, and 269 for the control. The percentage difference in favorable outcomes between the two groups would be 10.0 (p-value = 0.002) resulting in misclassification error of 2.1% (10.0–7.9).

Thus, corrections for all three patterns of misclassification demonstrated a potential for *greater* outcome differences and *smaller* p-values than the observed dataset if the study assumption was true and the misclassification existed in the observed outcome measurement.

Misclassification and Outcome Distribution

Table 2 illustrates the relationship between misclassification and the dichotomous outcome distribution under three misclassification models. A hypothetical 800-patient dataset (400 patients each group) with a 55% favorable outcome rate and 20% mortality rate was used for this illustration.

In general, without an outcome difference (i.e., 0% treatment effect), the misclassified outcome numbers were the same for both treatment and control groups and the misclassification only resulted in outcome distribution shifts, but not in outcome differences for all three models. However, with an outcome difference (i.e., 5%, 10%, 15% treatment effect), the outcome distributions for the treatment group were different from the distributions for the control group. As a result, the misclassified outcome numbers for the treatment and control groups were also different. For example, a 20% outcome number exchange between the favorable and unfavorable outcome categories in the random misclassification case caused the 5%, 10%, and 15% outcome differences to decrease to 3.2%, 6.4%, and 9.6%. The reduction is 1.8%, 3.6%, and 5.4%, respectively, from the previous outcome differences.

Similarly, in the *upward* (20% up and 10% down) and *downward* (20% down and 10% up) misclassification examples, after applying the rate exchange between the dichotomous outcome categories, the outcome differences (i.e., 5%, 10%, 15%) decreased to 3.7%, 7.4%, and 11.1% (Upward), as well as 3.6%, 7.2%, and 10.8% (Downward), which were 1.3%, 2.6%, and 3.9% (Upward), as well as 1.4%, 2.8%, and 4.2% (Downward) reductions from the original outcome differences, respectively.

Thus, it is conceivable that the impact from a misclassification on a dichotomous outcome measurement is not only related to the misclassification but also depends on the outcome distributions of the two study groups. Regardless of the random, upward and downward patterns of misclassifications, for a fixed rate of misclassification, the dichotomous outcome difference depends on the size of treatment effect or the difference in outcome distribution between the treatment and control groups. This is illustrated in Table 2, where all three misclassification examples have revealed that the more the treatment group differs from the control, the greater the impact of misclassification.

Table 2. Effect of Misclassification on the Binary GOS Outcome Distribution

		GOS after misclassifications																		
		0% Misclassification				*Random: 20% up and 20% down*					*Upward: 20% up and 10% down*					*Downward: 10% up and 20% down*				
	Simulated treatment effect (%)				*Outcome*				*Outcome*					*Outcome*					*Outcome*	
		D	*V/SD*	*MD/G*	*%Dif.*[a]	*D*	*V/SD*	*MD/G*	*%Dif.*	*%Ded.*[b]	*D*	*V/SD*	*MD/G*	*%Dif.*	*%Ded.*	*D*	*V/SD*	*MD/G*	*%Dif.*	*%Ded.*
Control		80	100	220		80	124	196			80	102	218			80	134	186		
Treatment	0%	80	100	220	0	80	124	196	0		80	102	218	0		80	134	186	0	
	5%	76	84	240	5	76	115	209	3.2	1.8	76	91	233	3.7	1.3	76	124	200	3.6	1.4
	10%	72	68	260	10	72	106	222	6.4	3.6	72	80	248	7.4	2.6	72	113	215	7.2	2.8
	15%	68	52	280	15	68	98	234	9.6	5.4	68	70	262	11.1	3.9	68	103	229	10.8	4.2

[a]Outcome difference (%).
[b]Deduction (%) from the expected outcome difference.

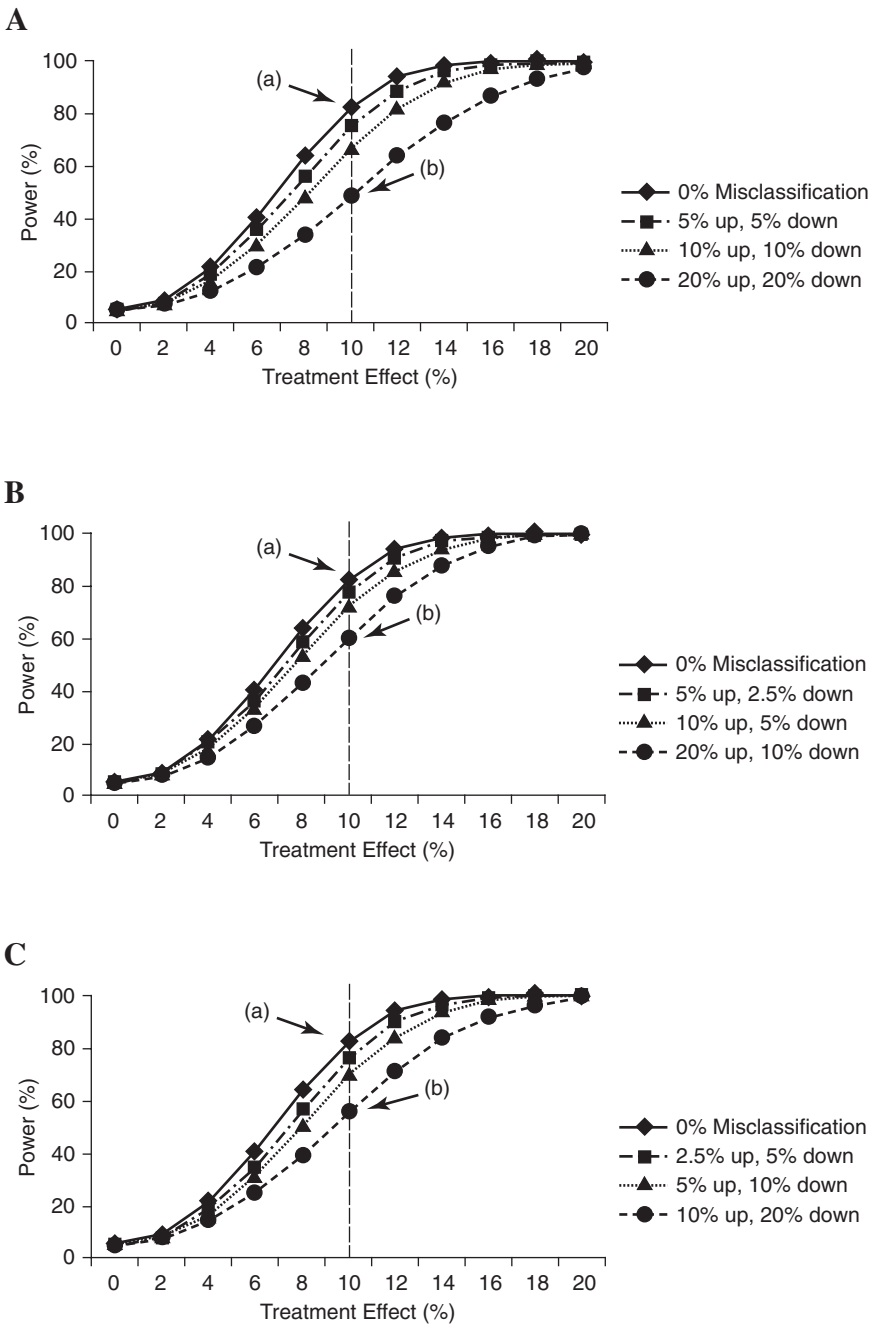

FIG. 1. (**A**) Effect of random misclassification pattern on the power. The solid line presents the correlation between the power and the expected treatment effect, and dashed lines present the correlation between the random misclassified treatment effect and power according to the symbol key. For example, for the case of a 10% treatment effect, a 20% up and down random misclassification would result in a reduction of power from 82% (point a) to 49% (point b) thereby rendering the trial non-significant. (**B**) Effect of upward misclassification pattern on the power. The solid line represents the correlation between the power and the expected treatment effect, and dashed lines represent the correlation between the upward misclassified treatment effect and power according to the symbol key. For example, for the case of a 10% treatment effect, a 20% up and 10% down (lowest dashed line) misclassification results in a reduction of power from 82% (point a) to 60% (point b). (**C**) Effect of downward misclassification pattern on the power. The solid line represents the correlation between the power and the expected treatment effect, and dashed lines represent the correlation between the downward misclassified treatment effect and power according to the symbol key. For example, for the case of a 10% treatment effect, a 10% up and 20% down (lowest dashed line) misclassification results in a reduction of power from 82% (point a) to 55% (point b).

Misclassification and Power

The powers of detecting the expected treatment effect and the misclassified treatment effect were compared and are illustrated in Figure 1 using the same hypothetical 800-patient dataset.

An example of the effect of random pattern on the power is shown in Figure 1a. Under a given treatment effect (i.e., the improved proportion of the favorable outcome in the treatment group), the power was inversely associated with the rate of the misclassification. For example, the power of detecting a 10% true treatment effect with a two-sided 95% significance is 82% (solid line); however, after the 5%, 10%, and 20% misclassifications were applied to the expected treatment effect, the power of detecting the same 10% treatment difference decreased to 75%, 67%, and 49%, respectively. Clearly, this was due to the altered outcome difference by the misclassification. The higher the misclassification rate, the smaller the treatment effect and the lower the power.

Figure 1b,c demonstrates the effect of the upward and downward misclassification patterns on the power. Similar results on reducing the power were observed, albeit in different degrees. If using a power to detect a 10% treatment effect with a two-sided 95% significance as an example, the upward pattern with a combination of 5% up and 2.5% down, 10% up and 5% down, and 20% up and 10% down were considered to be the rate exchange between the dichotomous outcome categories, then the desired 82% power would be decreased to 78%, 72%, and 60% accordingly. On the other hand, if the situation was reversed, namely downward pattern with 2.5% up and 5% down, 5% up and 10% down, and 10% up and 20% down were used as the rate exchanges between the outcome categories, then the expected 82% power would be reduced to 76%, 70%, and 55%, respectively.

DISCUSSION

Outcome Measurements and Outcome Misclassification in Trials of Head Injury

The GOS is widely used for TBI outcome measurement (Jennett and Bond, 1975) and recommended as primary endpoint for assessing efficacy of novel therapeutic approaches in clinical trials. For purposes of analysis in clinical trials, the GOS is commonly dichotomized into favorable versus unfavorable outcome, collapsing the five-point categorical outcome scale into a binary outcome measure (Bullock et al., 2002; Choi et al., 1998; Maas et al., 1997; Narayan et al., 2002; Teasdale et al., 1998; Wilson et al., 2002). Despite the acceptance of the GOS as a global functional outcome measure, it has been criticized as being insensitive, especially in the more favorable end of outcome (Bullock et al., 2002; Levin et al., 2001; Teasdale et al., 1998). The eight-point extended GOS (GOSE) has been introduced to increase sensitivity of outcome assessment, and the use of a structured interview is advocated to obtain more consistency in outcome assignment (Fayol et al., 2004; Wilson et al., 1998). Although the GOSE offers increased sensitivity, this benefit may be offset by a higher rate of misclassification. Recent evidence indicates an agreement rate as low as 60% in GOSE by untrained investigators (Wilson et al., 2007).

Misclassification, especially the nondifferential misclassification, is a relevant issue in clinical trial design. Previous work indicated that random misclassification could mask the true efficacy and reduce the power of finding a treatment effect (Choi et al., 2002). By understanding the consequence of outcome misclassification, efforts could be made to improve the accuracy and consistency of outcome measurements.

The present study has confirmed the substantial effects of nondifferential misclassification on outcome analysis and statistical power under various scenarios. One may question whether the effects of misclassification are substantial enough to be important. Clearly, from our analysis, we have found that a treatment effect may be reduced from 10% to 6.8% by a 20% random misclassification (i.e., 20% up and 20% down), which is more than sufficient to render a trial ineffective. The effect of misclassification on treatment effect is summarized in Table 3.

Moreover, the scenarios and the rates of misclassification investigated are not unrealistic to clinical practice. Marmarou (2001) conducted a study within the American Brain Injury Consortium to ascertain the reliability of the GOS rating and found an upward shift of 17.4% of severe patients to the moderate disability category. An upward shift of outcome assignment had been previously reported (Anderson et al., 1993) and is a likely result of the optimism of the patient's primary care providers who compare the improved outcome to the serious condition immediately after injury, rather than to the healthy pre-injury status. Conversely, a rigid application of the criteria from the structured interview or questionnaires by research workers tends to allocate patients to lower outcome categories (Teasdale et al., 1998; Wilson et al., 2007). Therefore, nondifferential misclassification may be found in either the upward or downward direction, based on different clinical scenarios. It is for this reason that three patterns were studied in this report.

An Assessment of Three Possible Misclassification Patterns

To demonstrate the effect of nondifferential misclassification on the outcome, we applied three possible patterns of misclassification on a real Phase III head injury

Table 3. Reduction of Treatment Effect and Power by Misclassification

Patterns of misclassification[a]	*Treatment effect reduction*	*Power reduction*[b]
Random		
10% up/down	10% → 8.4%	80% → 66.5%
20% up/down	10% → 6.8%	80% → 48.6%
Upward		
10% up	10% → 9.2%	80% → 76.6%
20% up	10% → 8.4%	80% → 69.9%

[a]Misclassification on both treatment and control arms, assume 55% favorable outcome and 20% mortality.

[b]Two-arm trial, $n = 800$; expected treatment effect = 10%; power = 80% and 95% two-sided significance.

trial data using five-category GOS outcome distribution in Table 1. According to the results from previous studies, we assume that there were certain rates of nondifferential misclassification embedded in the observed dataset, we corrected the hypothetical misclassified outcome numbers to the observed data using three models. After the numbers were corrected, a larger outcome difference and a smaller *p*-value were revealed in all three misclassification patterns.

Therefore, if the misclassification indeed existed in the past trial dataset as described in this study and as suggested by other studies, then the true outcome difference would have been larger. More importantly, our study indicated that regardless of which direction the dichotomous outcome was misclassified (i.e., random, upward, and downward), the effect of nondifferential misclassification always tends to reduce the true dichotomous outcome difference.

It should be noted that the random and the downward patterns in our examples seemed to have a larger effect on reducing the outcome difference than the upward pattern. This is likely due to the outcome distribution being misclassified and the rates of misclassification being applied. For example, more outcome numbers were exchanged from the category of MD/GR (i.e., 20% MD/GR = (0.2)*(252) = 50) with the numbers of V/SD in the random or downward cases, as compared with the numbers that were exchanged (10% MD/GR = (0.1)*(252) = 25) in the upward case. Thus, it is reasonable to understand why the random and the downward patterns had a larger impact on the outcome difference in our example.

In summary, the true outcome difference is always affected more by a higher misclassification rate and a larger difference in outcome distributions between the treatment and control groups. Therefore, any procedures that minimize the misclassification, such as proper outcome measurement techniques and the methods for improving the inter-observer agreement, should be indicated according to this study. Experience in the recent Phase III clinical trial on Dexanabinol showed that training of outcome assessors can be highly effective (Wilson et al., 2007).

Differential Effects of Misclassification in Treatment and Control Groups

Although it is generally assumed that the rate of the misclassification under a blinded clinical trial condition is the same for both control and treatment groups (i.e., nondifferential or random misclassification), the effect of the misclassification on these two are unlikely to be the same in the presence of a treatment effect. This is a consequence of the different outcome distribution between the treatment groups, as illustrated in Table 2.

For example, in the random misclassification (20% up/down) case, with no treatment effect, the misclassified outcome numbers are the same for both treatment and control groups; the misclassification only resulted in an outcome distribution shift but not in an outcome difference. However, with a treatment effect (i.e., 5%, 10%, 15%), the misclassified outcome numbers for the treatment and control groups are no longer the same, i.e., more patients' outcome in the treatment group are affected by the misclassification due to a larger outcome difference. Thus, instead of having an 5%, 10%, 15% in treatment effect, only a 3.4%, 6.8%, 10.2% outcome difference results, which represents a 1.6%, 3.2%, and 4.8% reduction of the previous outcome difference, respectively.

The other two patterns followed a similar trend as well. Figure 2 shows the example of correlation between upward misclassifications and reduction of treatment effect using same hypothetical 800-patient data as in Table 2. For a fixed rate of misclassification, the outcome difference depends on the size of treatment effect or the difference in outcome distribution between the treatment and control groups. For example, after a 20% upward misclassifica-

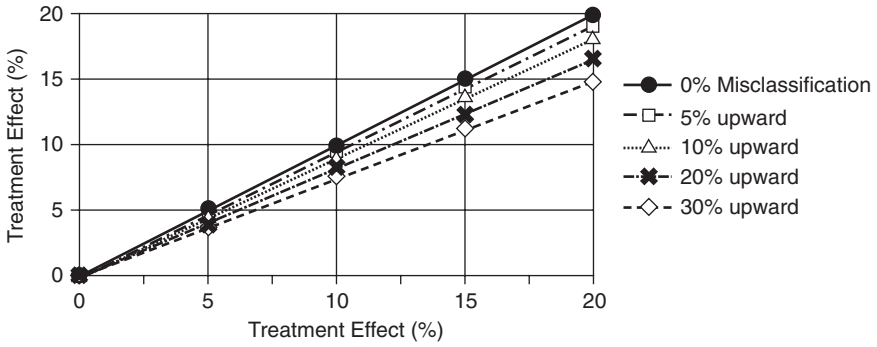

FIG. 2. Reduction of treatment effect by upward outcome misclassification. The solid line represents the expected treatment effect, and dashed lines represent the reduction of treatment effect by the upward misclassifications according to the symbol key. For example, for the case of a 10% treatment effect, a 30% up (lowest dashed line) misclassification results in a reduction of treatment effect to 7.6%, which is a 2.4% reduction from the expected 10% treatment effect.

tion, the expected 5%, 10%, 15%, and 20% outcome differences were reduced to 4.2%, 8.4%, 12.2%, and 16.6%, respectively. On the other hand, for a fixed treatment effect, the effect of misclassification on outcome difference depends on the rate of misclassification. For example, after 5%, 10%, 20%, and 30% biased upward misclassification, an expected 10% treatment effect was reduced to 9.4%, 9.2%, 8.4%, and 7.6%, respectively.

The implication here is that, if a study drug does have an effect on improving the patients' outcome, the treatment group is likely to be affected more by the misclassification than the control group. The more a treatment differs from the control, the greater the number of patients affected by the misclassification, leading to a greater reduction in the true outcome difference.

It is important to note that the demonstration on the dichotomous outcome distribution can also be applied to more than two category distributions. For example, if there is a larger difference between MD and GR, the true difference between these two categories will be affected more by misclassification. Likewise, if a larger difference exists between SD and MD, the actual difference between these two will be decreased more as a result of the misclassification. This topic will be studied in greater detail in the future.

Dealing with Misclassification

Since all GOS categories can be misclassified except death and as the affected outcome numbers are associated with the treatment effect and/or the outcome distribution, one might relate the issue to the choice of outcome measurements in head injury clinical trials. One study has suggested that an increase in outcome categories leads to an increase in misclassifications, and several other studies have proved that inter-observer disagreement is much higher in the eight-category GOS than that in the five-category GOS (Choi et al., 2002; Maas et al., 1983). These observations underline the notion that the outcome measurement with fewer outcome categories might be less affected by the misclassification. A careful balance will need to be sought between the desire for more sensitive expanded outcome measures and adverse effects of misclassification.

We suggest that both outcome misclassification and the sensible outcome measurement are important issues in the TBI trial design, which, in turn, is directly associated with the success of a trial. However, both the strategy to minimize the outcome misclassification, and to select a sensible outcome measurement should be considered separately. Although outcome misclassification is unavoidable, it is possible that errors in classification may be reduced. Accordingly, procedures such as structured interviews, proper outcome information resources, quality assurance of outcome evaluation and properly trained personnel have been previously shown to be successful approaches of minimizing the misclassification (Marmarou, 2001; Pettigrew et al., 2003; Wilson et al., 1998, 2007). These measures as well as developing new strategies are recommended in the clinical trial design.

On the other hand, carefully examining the outcome distribution from Phase II trials and selecting a sensitive outcome measurement to match each individual outcome distribution should be considered. For instance, if a treatment effect was mainly focused between the moderate and severe disability categories or other adjacent GOS categories, a dichotomous outcome would be a better choice over more GOS categories (Choi et al., 2002). However, if more or all categories of the GOS are affected by the treatment, then the dichotomous GOS would be less powerful than using more GOS categories

Effect of Misclassification on Power and Sample Size

Recognizing that outcome misclassification has a significant potential to reduce the true treatment effect, one would naturally relate this consequence to the power and sample size of a trial design. For a typical Phase III TBI trial, a sample size of 800 patients (i.e., 400 patients in treatment group, 400 in placebo group) is required in order to detect an absolute treatment effect that increases the proportion of favorable outcomes from 50% to 60%, with 80% power and 5% significance. We used a similar design to examine the effect of outcome misclassification on the desired power in the TBI trial. The correlation between the power and three potential patterns of misclassification was depicted in Figure 1.

As one might expect, in parallel with the effect of reducing the true outcome difference, all three patterns of misclassification have an inverse effect on the power. For instance (Fig. 1a), without misclassification, the expected power of detecting a 10% treatment effect (i.e., improving favorable outcome from 55% to 65% in the treatment group in our example) was 82%; with the same condition and a 10% random (i.e., 10% up and 10% down) misclassification for both study groups, the power of detecting such effect decreased to 67%; similarly, the powers under the upward (i.e., 10% up and 5% down), and the downward (i.e., 10% down and 5% up) misclassification condition reduced the power from 82% to 72% and 70%, respectively. Clearly, the examples shown in this study demonstrate that the desired power to detect the treatment effect could be compromised by misclassification of the dichotomous GOS outcome; the greater the number of outcomes misclassified, the greater the degree of power compromised.

Compensation for Reduced Power Due to Misclassification

As misclassification reduces power, it would seem reasonable to simply increase the sample size to compensate for the power reduction. This can be done. However, increasing the sample size can only raise the power but cannot compensate for treatment effect due to misclassification. Using our previous example in Table 3, a 10% random misclassification can reduce the original 10% treatment effect to 8.4%, and the power was subsequently reduced from 80% to 66.5% for detecting 8.4% treatment effect. In this example, one can increase the sample size from 800 to 1094 in order to raise the power from 66.5% to 80% for detecting 8.4% treatment effect, but still, the increased sample size cannot compensate the 1.6% (10%–8.4%) treatment reduction. This further emphasizes the importance of designing procedures to minimize the effect of misclassification. In summary, the only way to blunt the reduction of treatment effect is to reduce misclassification.

CONCLUSION

All three patterns of nondifferential misclassification act to attenuate the treatment effect and reduce the power of detecting the true treatment effect. In the case of a positive drug effect, misclassification leads to a conservative estimation of the true efficacy. The magnitude of such influence not only depends on the size of the misclassification, but also on the magnitude of treatment effect. Nondifferential misclassification directly reduces the power of finding the true treatment effect. If the outcome of the treatment arm is worse, then misclassification acts to blunt the difference between placebo and treatment. Thus, an awareness of this procedural error and methods to reduce misclassification should be incorporated in TBI clinical trials.

ACKNOWLEDGMENTS

Grant support was provided by NS-042691 and NS019235-21.

AUTHOR DISCLOSURE STATEMENT

No conflicting financial interests exist.

REFERENCES

Anderson, S.I., Housley, A.M., Jones, P.A., Slattery, J., and Miller, J.D. (1993). Glasgow Outcome Scale: an inter-rater reliability study. Brain Inj. **7,** 309–317.

Bolland, K., Sooriyarachchi, M.R., and Whitehead, J. (1998). Sample size review in a head injury trial with ordered categorical responses. Statist. Med. **17,** 2835–2847.

Brooks, D.N., Hosie, J., Bond, M.R., Jennett, B., and Aughton, M. (1986). Cognitive sequelae of severe head injury in relation to the Glasgow Outcome Scale. J. Neurol. Neurosurg. Psychiatry **49,** 549–553.

Bullock, M.R., Merchant, R.E., Choi, S.C., Gilman, C.B., Kreutzer, J.S., Marmarou, A., and Teasdale, G.M. (2002). Outcome measures for clinical trials in neurotrauma. Neurosurg. Focus **13,** ECP1.

Choi, S.C., Clifton, G.L., Marmarou, A., and Miller, E.R. (2002). Misclassification and treatment effect on primary outcome measures in clinical trials of severe neurotrauma. J. Neurotrauma **19,** 17–22.

Choi, S.C., Marmarou, A., Bullock, R., Nichols, J.S., Wei, X., and Pitts, L.H. (1998). Primary end points in phase III clinical trials of severe head trauma: DRS versus GOS. The American Brain Injury Consortium Study Group. J. Neurotrauma **15,** 771–776.

Fayol, P., Carriere, H., Habonimana, D., Preux, P.M., and Dumond, J.J. (2004). French version of structured interviews for the Glasgow Outcome Scale: guidelines and first studies of validation. Ann. Readapt. Med. Phys. **47,** 142–156.

Hukkelhoven, C.W., Steyerberg, E.W., Farace, E., Habbema, J.D., Marshall, L.F., and Maas, A.I. (2002). Regional differences in patient characteristics, case management, and outcomes in traumatic brain injury: experience from the tirilazad trials. J. Neurosurg. **97,** 549–557.

Jennett, B., and Bond, M. (1975). Assessment of outcome after severe brain damage. Lancet **1,** 480–484.

Levin, H.S., Boake, C., Song, J., McCauley, S., Contant, C., Diaz-Marchan, P., Brundage, S., Goodman, H., and Kotrla, K.J. (2001). Validity and sensitivity to change of the extended Glasgow Outcome Scale in mild to moderate traumatic brain injury. J. Neurotrauma **18,** 575–584.

Maas, A.I., Braakman, R., Schouten, H.J., Minderhoud, J.M., and van Zomeren, A.H. (1983). Agreement between physicians on assessment of outcome following severe head injury. J. Neurosurg. **58,** 321–325.

Maas, A.I., Dearden, M., Teasdale, G.M., Braakman, R., Cohadon, F., Iannotti, F., Karimi, A., Lapierre, F., Murray, G., Ohman, J., Persson, L., Servadei, F., Stocchetti, N., and Unterberg, A. (1997). EBIC-guidelines for management of severe head injury in adults. European Brain Injury Consortium. Acta Neurochir. **139,** 286–294.

Marmarou, A. (2001). *Head Trauma: Basic, Preclinical, Clinical Direction*, 1st ed. Wiley: New York.

Narayan, R.K., Michel, M.E., Ansell, B., Baethmann, A., Biegon, A., Bracken, M.B., Bullock, M.R., Choi, S.C., Clifton, G.L., Contant, C.F., Coplin, W.M., Dietrich, W.D., Ghajar, J., Grady, S.M., Grossman, R.G., Hall, E.D., Heetderks, W., Hovda, D.A., Jallo, J., Katz, R.L., Knoller, N., Kochanek, P.M., Maas, A.I., Majde, J., Marion, D.W., Marmarou, A., Marshall, L.F., McIntosh, T.K., Miller, E., Mohberg, N., Muizelaar, J.P., Pitts, L.H., Quinn, P., Riesenfeld, G., Robertson, C.S., Strauss, K.I., Teasdale, G., Temkin, N., Tuma, R., Wade, C., Walker, M.D., Weinrich, M., Whyte, J., Wilberger, J., Young, A.B., and Yurkewicz, L. (2002). Clinical trials in head injury. J. Neurotrauma **19,** 503–557.

Pettigrew, L.E., Wilson, J.T., and Teasdale, G.M. (2003). Reliability of ratings on the Glasgow Outcome Scales from in-person and telephone structured interviews. J. Head Trauma Rehabil. **18,** 252–258.

Scheibel, R.S., Levin, H.S., and Clifton, G.L. (1998). Completion rates and feasibility of outcome measures: experience in a multicenter clinical trial of systemic hypothermia for severe head injury. J. Neurotrauma **15,** 685–692.

Teasdale, G.M., Pettigrew, L.E., Wilson, J.T., Murray, G., and Jennett, B. (1998). Analyzing outcome of treatment of severe head injury: a review and update on advancing the use of the Glasgow Outcome Scale. J. Neurotrauma **15,** 587–597.

Wilson, J.T., Edwards, P., Fiddes, H., Stewart, E., and Teasdale, G.M. (2002). Reliability of postal questionnaires for the Glasgow Outcome Scale. J. Neurotrauma **19,** 999–1005.

Wilson, J.T., Pettigrew, L.E., and Teasdale, G.M. (1998). Structured interviews for the Glasgow Outcome Scale and the extended Glasgow Outcome Scale: guidelines for their use. J. Neurotrauma **15,** 573–585.

Wilson, J.T., Slieker, F.J., Legrand, V., Murray, G., Stocchetti, N., and Maas, A.I. (2007). Observer variation in the assessment of outcome in traumatic brain injury: experience from a multicenter, international randomized clinical trial. Neurosurgery **61,** 123–129.

Address reprint requests to:
Anthony Marmarou, Ph.D.
Department of Neurosurgery
Virginia Commonwealth University
P.O. Box 980508
Richmond, VA 23298-0508

E-mail: marmarou@vcu.edu